Practice on assessing grammar and vocabulary: The case of the TOEFL

ZHUANG Xin

(College of Foreign Languages, Zhejiang Gongshang University, Hangzhou Zhejiang 310018, China)

Abstract: The Test of English as a Foreign Language (TOEFL) brings tremendous influence to EFL (English as a Foreign Language) learners worldwide. TOEFL 2000 project claims that TOEFL, as a more reflective of communicative model, could provide more information about international students' language ability that it is supposed to measure. However, after detailed analyzing an authentic paper-based test paper in May, 2001 in China as a sample from four aspects—test reliability, construct validity, authenticity and interactiveness respectively, it is found that the test puts too much emphasis on vocabulary and grammar knowledge within almost every session of the test paper, in which "structure and written expression" could be the most disputed part. The content could not fully demonstrate its validity and communicative purposes so that it is suspected that test takers could meet the later demands in academic study abroad. Nevertheless, this is a powerful explanation about the current revolutionized change in the framework and content of TOEFL to meet the principles of designing a test, which could provide more information and guidance for later test designs.

Key words: TOEFL; validity; grammar; vocabulary; communicative purposes

During the couple of years, TOEFL (Test of English as a Foreign Language) has undergone a revolutionized change in the test content and framework. What makes the change? What are the changes? What are the implications in the changes? Answering these three W-questions could provide us a guideline for making language tests much more reliable, valid, authentic and interactive in accordance with the communicative language teaching worldwide.

1. Background knowledge about the TOEFL

The eagerness of learning a foreign language promotes the development of foreign language learning. In order to prove one's language proficiency, the TOEFL, as one form of international language tests, becomes the dominant type worldwide. It is slightly different from tests in the classroom. It has no fixed content that have been taught to test takers, which decides its wide range and general contents towards EFL learners worldwide. It is rather a proficiency test than an achievement test since it measures someone's language abilities at a certain time. The TOEFL test is norm-referenced test but not criterion-referenced one since test results are interpreted with reference to the performance of a certain group, whose performance is used to relates one candidate's performance to that of other candidates (Hughes, 1989, pp. 17-18), that is, to obtain meaning from the referenced scores (Ebel & Frisbie, 1991, p. 34).

TOEFL 2000 project claims that TOEFL is "more reflective of communicative competence models" and it "provides more information than current TOEFL scores do about international students' ability to use English in

ZHUANG Xin, lecturer, College of Foreign Languages, Zhejiang Gongshang University; research fields: English language teaching, teacher education.

an academic environment" (Jamieson, et al., 2000, p. 3). Before the birth of the TOEFL 2000 project, some researchers categorized TOEFL as a non-communicative test. But does it really make a revolutionized change? As Morrow (1986, p. 9) mentions that in communicative testing, "What we are concerned with is the performance of an individual performing a set of tasks in a foreign language". Can it really attain its ambitious goals?

According to the TOEFL 2000 project, the traditional TOEFL test exams one's language competence in listening, reading and writing skills, among which integrated with vocabulary and structure knowledge for the years around. Moreover, there are standard procedures for administering and scoring the test and TOEFL that is held systematically in fixed work-based worldwide and the total paper-based test score is now reported on a scale that ranges from 310-677, while TWE (Test of Written English) score is reported separately on a scale of 1-6. Finally, through a process of empirical research and development, the characteristics of the tests are well-known, and the testees even have suggestions and tips of preparing a TOEFL test, which are provided by Educational Testing Service (ETS). In the survey done by Brown and Ross (1996, p. 233), there are approximately 85.2% testers using the TOEFL test score for graduate, undergraduate studies or another type of school, 13.8% ones for a license or a company and only 1% people give no reason for taking the TOEFL tests among 20,000 randomly selected testees. Evidently, more and more people use TOEFL score as a proof to demonstrate the individual language proficiency to meet the later requirements from both academic degree programmes and ESL learning as well, even though there is no standard criterion to define which score is a "pass" and which is a "failure".

As a large scale proficiency test, TOEFL is designed to measure people's language abilities. However, it is not a test to discover whether someone has adequate command of the language for a particular purpose but rather the one with more general concept. It is a common sense that TOEFL has been thrived for a long period to meet the global requirements on EFL testing due to either its rationality or its exclusiveness, but definitely it is meeting the new challenges from other test systems as the time goes by. For instance, more and more countries, especially the European countries adopted International English Language Testing System (IELTS) as a main assessment of English proficiency. This is not the national preference makes the tendency but undoubtedly reflects the basic considerations and appealing that come from the test principles.

2. Study on the TOEFL paper 2001

In order to have better understanding about some revolutionized changes of TOEFL in recent years, it is sensible to have a review on its tests based on TOEFL 2000 project.

2.1 Test framework

Take one TOEFL paper-based test for example, it was taken in May, 2001 in China generally. The whole structure of the test paper mainly consists of four parts:

(1) Section 1: Test of Written English (TWE) (30 minutes);

(2) Section 2: Listening Comprehension (30 minutes);

(3) Section 3: Structure and Written Expression (25 minutes);

(4) Section 4: Reading Comprehension (45 minutes).

Among the sections, section 2, 3 and 4 are timed tests in multiple-choice format with four options for each question. TOEFL, as a popular norm-referenced test for the whole world, is designed not based on certain contents or a language course but to meet the fundamental and necessary requirement of using language—to communicate. As the foremost aspect in the criteria, thus, we have to be careful about the designing and to reconsider the

function of the tests. To measure language proficiency in almost every aspect of situations, we need to take account of when, where, how, why and what is to be used. Therefore, how would the tests be as representative as possible is the key issue in designing language tests. Bachman and Palmer (1996, pp. 19-25) provide us some basic criteria which need to be reflected in the test paper. They are test reliability, construct validity, authenticity and interactiveness.

2.2 Test reliability

The concept of reliability is particularly important in the language tests. Although we can never have complete trust in any set of the scores, we try to produce a perfect and consistent test score which is free from measurement error mainly intrigued by different testing times, test forms, raters and other characteristics of the measurement context, that is, to concern the consistency of test judgements and results (Bachman, 1990; Hughes, 1989; Weir, 1990; Davies, 1990). And the highly reliable score ought to be “accurate, reproducible and generalizable to other testing occasions and other similar test instruments” (Ebel & Frisbie, 1991, p. 76). In TOEFL, there are two components of test reliability we need to consider, one is the performance of testers and the other is the reliability of the scoring. Let’s look at the data provided by ETS diachronically. In China, there were 31,462 students took TOEFL CBT between July 1999 and June 2000, in which the average scores in three parts listening, structure and reading were 20, 21, 21 respectively and the mean of total score was 206. Between July, 2001 and June, 2002 there were 58,772 students took TOEFL CBT, and they got 20, 21, 21 separately in three parts, the mean of total score was 207 (TOEFL test score and data 2000-2001, 2002-2003). From the data above, we could find that the scores of Chinese students generally cluster around the 20 level and the reliability estimates were well within the desirable range and substantial. Part of the reason is that the mark of TWE does not add into the whole score so that other three sessions require no judgement on scoring for the testing format, and could be in practice carried out by a computer, thus, the main part of TOEFL test is said to be objective and highly reliable.

2.3 Test validity

It seems to be axiomatic that “validity cannot be established unless reliability is also established for specific contexts of language performance” (Cumming & Mellow, 1995, p. 77). “A test, part of a test, or a testing technique is said to have construct validity if it can be demonstrated that it measures just the ability which it is supposed to measure” (Hughes, 1989, p. 26). If test scores are affected by other abilities rather than the one we want to measure, they will not be the satisfactory interpretation of the particular ability. In this TOEFL test paper of May 2001, if we look at each session rather than the holistic structure, reading comprehension won’t cause too much concern since it is fairly demonstrate, which measures a distinct ability. There are five pieces of articles, related with social science, biology, literature, ethology and geology, which covered wide varieties of topics. Including these fifty questions, the whole reading comprehension has 3,673 words, which means that the testees need to finish reading in about 82 words per minute. This is a high demand for EFL learners who need to prove their abilities in language knowledge as well as cultural background knowledge. What’s more the reading part not only questions the related information but also questions the implied meaning and even the specific meaning of a certain word. From those aspects we need the skills of reading both extensively and intensively. If “the purpose, events, skills, functions, levels are carried out as what they are expected to” (Carroll, 1980, p. 67), the construct validation is fully displayed in the TOEFL reading part.

2.4 Test authenticity and interactiveness

The other two principles we need to concern are authenticity and interactiveness. “Authenticity provides a means for investigating the extent to which score interpretations generalize beyond performance on the test to

language use" (Bachman & Palmer, 1996, pp. 23-24), which means the task that the test set is correspond with the content of the test. In the language test, authenticity sometimes distantly related with real communicative tasks by carrying out series of linguistic skills rather than genuine operational ones for reliability and economy (Carroll, 1980, p. 37). The listening comprehension in TOEFL test simulates the speaking environment in the North American colleges or universities and adds some idiomatic expressions common to spoken English to attain the features of the target language usage, which we could say this session provides the authentic materials in a certain extent. Nevertheless, for the language proficiency if we only test listening or reading, the whole test are not fully activated and we would never have the generalized idea about the testees' language standard so that the test could not be called successful at all.

Interactiveness refers to the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task (Bachman & Palmer, 1996, p. 25). Due to the different areas of language knowledge, planning strategy and personality, how could we give each testee a fair chance is always a question. TOEFL test demonstrates this point by offering a general topic in writing, by providing standard written English in grammar structure, and by covering various topics in reading, however, we still could find something which is too "Americanized". For instance, the pronunciation of listening comprehension is sounded in American way which seems to be a hard work for the learners whose first language is not English worldwide.

Compared with the TOEFL listening section, Cambridge First Certificate in English (FCE) provides a variety of accents in both standard variants of English native speaker accent and English non-native speaker accents (Cambridge FCE Handbook, 1997). These designs in FCE initiate the similar environment in English countries and make the whole test more communicative and practical. Many articles of reading comprehension concern lots of American topics but fairly rare non-American ones, although it seems to cover abundant topics. As Hilke and Wadden (1997, p. 36) note that "what certain TOEFL texts choose to include, moreover, is often as significant as what they fail to include". In this test paper, two fifths of the reading contents attach closely with American history background. Thus, whether the TOEFL test provides each candidate a fair chance is not clearly demonstrated.

3. Analyzing the "language" knowledge in the TOEFL

In the framework of the language structure put forward by Bachman and Palmer (1996, pp. 68-75), we could infer that learners' language ability consists of two parts, one is language knowledge and the other is strategic competence/metacognitive strategy. That is to say, learners need to know the vocabulary, grammar, sound system as well as to use the coherent sentences in a certain language setting to achieve the communicative goals of language users. The TOEFL test, the way to demonstrate candidates' achievement in English, should determine whether they could apply the knowledge and skills in their future real-life study, that is, to assess their performance in this language. This is the main reason to construct the tests to get the information: "How well individuals perform on the test represents to some degree how they might be expected to respond outside the testing environment" (Sax, 1997, p. 304). However, we do not expect a test can measure all the aspects of language in each section, thus, the samplings should be as represented as possible. And here more emphasis will be put on the grammatical knowledge part in the TOEFL test.

3.1 Testing grammatical knowledge in writing, listening and reading skills

Grammatical knowledge mainly includes three parts: vocabulary, syntax and phonology (Bachman & Palmer,

1996, p. 70). In this TOEFL test paper, knowledge of vocabulary seems to be tested in all the sections, which proves the common sense that words are basic building blocks of language. Vocabulary, which is embedded, comprehensive and context dependent in nature, plays an explicit role in the assessment of learners' performance (Read & Chapelle, 2001). The best way to test people's vocabulary is to use various ways to test either the basic meaning of a word, or its derived form, its collocations or its meaning relationship in a context. Nation (1990) gives a systematic list of competencies which has come to know as types of word knowledge, which are (1) spoken form of the word; (2) written form of the word; (3) grammatical behaviour of the word; (4) collocational behaviour of the word; (5) frequency of the word; (6) stylistic register constrains of the word; (7) conceptual meaning of the word; (8) associations the word has with other related words (Schrutt, 1999, p. 194). These word knowledge types decide the meaning of knowing a word, thus, if we want to analyze the construct validity of vocabulary items in TOEFL, whether the meaning sense is typical way of usage in an academic context in the future is the key element. Schrutt (1999, p. 192) also points out: "Although any individual vocabulary item is likely to have internal content validity, there are broader issues involving the representativeness of the target words chosen".

In TWE it checks not only the written form of the words but also the function and collocations of their grammatical usage. Cumming and Mellow (1995, p. 77) define a general ESL composition profile, which is "vocabulary (range, choice, usage, word form mastery, register), language use (complex constructions, errors of agreement, tense, number, word order/function, articles, pronouns, prepositions) and mechanics (spelling, punctuation, capitalization, paragraphing)". The testees need to finish a composition in 30 minutes which is constituted by more than 300 words are more preferable. However, the limitation in TWE is its limited styles of writing. Like the topic in this test paper, most of the writing style in the TOEFL is a contrastive writing to show personal preference or the choice. Although the writing section is not the specific part to test grammatical knowledge, whether the sample chosen in the TOEFL test is truly the representative of the communicative competence is still a question.

In listening comprehension, testing vocabulary is not limited to single word any more. There are many compound words, phrases and even idiomatic expressions and slang. For example, in May, 2001 test paper, there are some idioms in the dialogues between two speakers like "*have something checked out, headed one's way, big show storm, get a little carried away, that sure beats sticking around here*" etc. Since most of the dialogues are selected from American daily life, lots of phrases and sentences cause great difficulty for EFL testees since it is difficult to work out the meaning by the surface meaning of the words. Moreover, both the conversation and the choices have a high demanding on grammar to require the testers give definite response in fifteen seconds. For example, four choices in No. 8 display four different tenses: the present simple, the past perfect, the subjunctive mood in future sense and the future tense. And the dialogue in No. 8 is:

> M: My back has been aching ever since I started playing tennis on the weekends.
> W: Haven't you had that checked out yet?
> Q: What does the woman imply?

From this short dialogue, we notice that usually the first person present the content or the background of their conversation and the second person gives the hint to the answer of the question. Summing up the questions from first thirty short dialogues, we get the following results (Table 1):

Table 1 Questions of 30 dialogues in the TOEFL test, May 2001

Typical questions	Percentage
What does the man/woman imply?	37%
What does the man/women mean?	33%
What does the man/woman suggest	13%
What can be inferred from the conversation?	10%
Others	7%

From the type of questions, it is not so difficult to find out that answering these questions in "listening comprehension" needs either fluency and consolidated grammatical knowledge. Listening comprehension test is much more a combination of testing on both vocabulary and syntax.

The communicative philosophy of reading test is to test "in what situations do we read which texts for which purposes" (Wijgh, 1995, p. 155). Originally, TOEFL tests had vocabulary items, which were selective and context-independent multiple-choice items presenting words in isolation. They were criticized since international students simply spent time unproductively memorizing long list of words together with synonyms or definitions (Read & Chapelle, 2001, p. 14). And now the prominent feature in vocabulary items still exists in the TOEFL reading comprehension subtest, that is, the testing on the meaning of words or short phrases. Banerjee and Clapham (2003, p. 116) point out that although the previous section in the TOEFL test called the reading and vocabulary and now it is renamed as reading comprehension, which still consists two distinct tests: Reading and vocabulary. In this test paper, there are 20 questions related to the close meaning or referring meaning of the words or phrases, in which 16 of them are questions about words. These questions take up two fifths of the overall reading questions, and the second article has the largest number of questions, which are five in ten. These questions always demonstrate in several fixed way: "*The word 'lured' in line 19 is closest in meaning to…; the word "them" in line 11 refers to…*". Although the "closest in meaning to" questions concern much of the word meaning in the context, the rest of the word questions seem to assess the range of candidates' vocabulary. And sometimes without referring back to the contents, testees still could get the answer if they simply know the meaning of words. As what Read (1997, 2000, cited by Read & Chapelle, 2001) has said those vocabulary items in the reading test of the TOEFL can be categorized into the relative independent group, despite the manner in which they are presented. Is it another section which focuses on the vocabulary again?

3.2 Testing grammar knowledge independently

Assessing language knowledge is always reflected in the "four-basic skills", speaking, listening, reading and writing. But considering some well-known proficiency tests erase the grammar component (Hughes, 1989, p. 141), the "structure and written expression" still remains as one part of the TOEFL tests, whose contents are similar to the section "use of English" in the First Certificate in English (FCE) in Cambridge Level Three. Wall, et al (1991, p. 214) suggest that if we want to decide the content validity, several elements need to be determined, that is, whether the tasks they are testing are the ones they intend to test; whether the sampling of tasks is adequate; and whether the level of difficulty of its components is proper. The principles for communicative language learning guiding test construction in "structure and written expression" suggest that testees should know how to use different structures and useful expressions in language output to be effective and efficient speakers and writers, which could satisfy the original purposes of studying in North America. In FCE, testers are expected to

demonstrate their knowledge on vocabulary and grammar and control of the language system[1] (First Certificate in English Handbook, 1997, p. 7).

Nevertheless, the TOEFL test gives a higher requirement on vocabulary and syntax, which both belong to the grammatical knowledge, by setting an independent section as a grammar-like test. There are two parts in the session: (1) incomplete sentences, with words or phrases as options (15 items); and (2) sentences in which some words or phrases are underlined. In the second part (25 items), the examinee must identify the words or phrases in each sentence that are not appropriate to standard, formal written English (Stevenson, 1987, p. 80). The forty items in this subtest need to be finished in 25 minutes which have a high demand on the speed of testing. In FCE, there are five parts needed to be finished in one hour and fifteen minutes, which are (1) multiple choice cloze; (2) open cloze; (3) "key" word transformations; (4) error correction and (5) word formation. (1) and (5) emphasize vocabulary while (4) focuses on grammar, and (2) and (3) concern the integration of grammar and vocabulary[2] (First Certificate in English Handbook, 1997, p. 28). There are 65 items in five sections of the FCE. Moreover, compared with the TOEFL test paper, it is evidently to see that the FCE has more variation on the form of the items and except part three which is produced on sentence-based questions, the other parts are all integrated with context. "Structure and written expression" in the TOEFL underwent no significant change in the recent revision of the test, and grammatical words are still the necessary parts in these exercises. Words like articles, prepositions, pronouns, conjunctions, auxiliaries, etc. are often referred to as *function words* which belong more to the grammar (Read, 2000, p. 18), as in the following example:

The hamster's basic diet is vegetarian, some hamsters also eat insects.
(A) Despite
(B) Although
(C) Regardless of
(D) Consequently

C*ontent words* such as nouns, "full" verbs, adjectives and adverbs provide links within sentences or modify the meaning of the content words (Read 2000, p. 18). Furthermore, there are many phrases become one of the testing focuses, as follows:

The giant ragweed, or buffalo weed, grows ____
(A) 18 feet up to high
(B) To high 18 feet up
(C) Up to 18 feet high
(D) 18 feet high up to

However, it is more appropriate to say that "structure and written expression" section in the TOEFL test is more like a grammar subtest rather than a simple vocabulary test, which has been considered "an important trait in the measurement of an individual's overall performance in a language" (Rea-Dickins, 1991, p. 115). The grammar subtest here cannot be regarded as a skill subtest like listening or reading subtests, so we cannot help but wonder "What does this grammar subtest measure? Is it a communicative test". When we talk about the communicative competence, often we concern with "generalized abilities" (Skehan, 1991, p. 9), the abilities to express one's

[1] Details of the TOEFL test introduction could be found at http://www.toefl.org/.
[2] Cambridge Examinations. (1997). *Certificates and Diplomas: FCE Handbooks*. University of Cambridge Local Examinations Syndicate.

meanings by using appropriate language in various contexts. In order to carry out a persuasive and rigorous assessment in the communicative test like TOEFL, we need to ensure that “the sample of communicative language ability in our tests is as representative as possible” (Weir, 1990, p. 11). Rea-Dickins (1991, p. 125) defines five factors to the “communicative” nature of a grammar test which I think two of them are quite necessary for the test: one is the contextualization of test items and the other is to put the instructions into focusing on meaning but not simply on form.

Before we look at the content validity of “structure and written expression” section, some elements are needed consideration. They are mainly the format of test item to be used, the area of the content to be sampled, the number of the items in the area, and the level of item difficulty (Osterlind, 1998, p. 78). The purpose on the requirement of contextualized test items is to meet the heuristic functions of knowledge. Providing authentic material to solve problems and to develop thinking is “highly relevant to communication in the discipline or occupation concerned”. And “the aim is to assess communicative proficiency in the subject concerned, not to test specific knowledge of it” (Carroll, 1980, p. 38). In both part one and part two of “structure and written expression” section of TOEFL test, testees mainly focus on the selection of the appropriate form on sentence-based format, which hardly have to exchange the information during the whole process of production, although most of the topics relate with academic areas. The limited items of testing grammatical knowledge could not reflect the testees’ knowledge of grammar completely and have no practical usage to prove that the testees would be competent enough to meet the later requirements on academic learning. Moreover, the written expression part in this section seems to test the writing ability in an indirect way. However, we suspect this section could really decide whether the learners’ competence in writing could meet the later demands in academic study.

One of the remarkable features in the TOEFL test is its format of answering the questions in four answer options—multiple choices. From the aspect of scoring, it is highly reliable since all the scoring could be carried out by computers, thus easily “discriminate between high- and low-achieving students” (Ebel & Frisbie, 1991, p. 124). And it could offer more flexibility for assessing a diversity of content and psychological processes (Osterlind, 1998, p. 163). For a four-option test, students have 25% point for each choice which seems to be more complex and less ambiguous than true-false decision item when making the correct choice. However, from the other side, not all the grammatical knowledge could be tested simply by choosing one correct answer from four choices. What’s more, the distractors of the test items decide a lot on the validity of the content. The choices could not attain the effect of checking certain knowledge if the other three choices have little relation with knowledge that wanted to test or without looking at the given context, no higher order thinking skills are needed. What’s more, students get credit to recognizing the wrong options through a process of elimination or simply guessing even though they still could not identify the correct option (Sax, 1997, p. 106; Ebel & Frisbie, 1991, p. 156). Moreover, the weakness of setting up easily-devised and objectively-scored tests of strings of linguistic items is that the testees may miss the essence of the measurement of communicative performance (Carroll, 1980, p. 9) and inhibit creativity and original thinking, and reduce all important knowledge to superficial facts (Osterlind, 1998, p. 164). Therefore, when the grammatical knowledge is simply tested in four-option multiple-choice, two criteria have been presented either to the task in the test item or the test takers’ performance.

In order to have a better understanding of the contents of “structure and written expression” section, I borrowed the types analysis from Hilke and Wadden (1997, pp. 30-34) to give a brief generalization of this test paper.

Part A: Structure (Fill in the blank):

(1) WIAS (what Is A Sentence): about 47.7%;

(2) Word choice: about 27.7%;

(3) Word order: about 13.3%;

(4) Verb form: about 13.3%.

(WIAS is a category to show that one clause contains one subject and one verb. Word choice tests how to use appropriate words and phrases. Word order is to check the proper order of the words. Verb form concerns with tense or aspect.)

Part B: Written expression (error analysis):

(1) Part of speech error: 24%;

(2) Prepositional error: 16%;

(3) Verb form error: 16%;

(4) Plural: 8%;

(5) Pronoun error: 8%;

(6) Redundancy error: 8%;

(7) Word order: 8%;

(8) Article error: 8%;

(9) Conjunction error: 4%.

From the data above, seventy-five percent of all TOEFL questions in part A are categorized into the WIAS (What Is A Sentence) and word choice errors groups. In part B more than half of the errors (56%) belong to three major types: part of speech, preposition and verb form. Grammar knowledge points in this test paper are relatively non-complex and less various. From the analysis above, we could find that some knowledge points in the items are repeatedly tested. If students are reinforced these basic structural rules, they will quickly improve their accuracy by answering these questions since grammar knowledge points in the TOEFL test are incomplete and limited. Furthermore, although the multiple-choice test of grammatical knowledge in “structure and written expression” could produce consistent or reliable scores, it is not sufficient to use this section as a placement subtest for writing. As Bachman and Palmer (1996, p. 23) noted that “grammatical knowledge is only one aspect of the ability to use language to perform academic writing tasks” since the area of language knowledge is quite narrow. However, high proportions of these sentences are compound and complex ones, which most of them are selected from academic articles, such as natural science, biography or history etc. with a large amount of professional and abstract words. Nevertheless, those complicated and academic words appear to be a superficial “threat” to the testers. Since one has a solid knowledge of grammar and notices the cohesion within a sentence, getting the answer is still an easy task. Like Bachman, et al (1995, p. 123) claimed that the TOEFL vocabulary items were judged to be more familiar for unprepared test takers. In sum, a multiple-choice format in a test has high reliability since it provides a relatively consistent result, but “reliability is not a sufficient condition for either construct validity or usefulness” (Bachman & Palmer, 1996, p. 23).

4. Implications

When thinking about the reliability of “structure and written expression” section in the TOEFL tests, the test developers want to set the minimum acceptable level of reliability as high as possible. Bachman and Palmer (1996, p. 135) provide two criteria to evaluate: One is “the way the construct has been defined”, the other is “the nature

of the test tasks". That is to say, only the test focuses on a relatively narrow range of components of language ability with relatively uniformed test tasks, could the test achieves higher levels of reliability. In "structure and written expression" section, it only adopts two types of tasks: One is to complete a sentence by selecting one of the four choices, the other is to choose one error from four underlined choices. They are all in the format of multiple-choice, but they demonstrate limited language abilities, which include language knowledge and strategic competence mentioned before. Therefore, with incomplete language knowledge and less variations in test task characteristic, the result is consistency, which is the essence of reliability.

What's the usage of the "structure and written expression" section? Can it really improve the English language proficiency as a part of communicative test? As Vollmer (1981, p. 154) mentioned, "most multiple-choice items are discrete-point items", but these items in "structure and written expression" differ from the ones in listening and writing in the TOEFL test. From the aspect of the contextualization of test items, listening and reading parts involve much more holistic understanding of the whole text, which integrate not only the abilities of listening or reading intensively or extensively, but the overall aspects of the communicative competence as well. In language proficiency testing, Read (1993, p. 357) recommended that words need to be understood in connected written or spoken discourse rather than just isolated items, which is very important to the EFL learners. In grammar test, the content should be defined more broadly than "syntax and morphology" and includes textual competence as well (Rea-Dickins, 1997, p. 92). The forms of the grammar test should be varied as to meet the original purpose of communicative testing, and some suggested techniques on like paraphrase, completion or gap-filling, guided short answer and summary and modified cloze could help to verify the test formats (Hughes, 1989, p. 143; Rea-Dickins, 1997, p. 91). Only when the grammatical knowledge integrates with language skills or it could be existed independently in variety of forms on either sentence-based or text-based format, could it be meaningful in the TOEFL tests. Furthermore, much more emphasis needs to be put on the meaning and communicative functions which presents its features on written expression rather than on the structure or form in the "structure and written expression".

Another issue of the "structure and written expression" in the TOEFL test is that how to design the proper choices in the multiple choices to make it more reliable to achieve its communicative goal. Thus, keeping a good balance between content words with useful expressions and functional words with structures reflects the main focus of this testing item. It is unwise to centralize too much on limited words but can still get the answer even neglecting the whole sentence. Moreover, like Huhta and Randell (1995, p. 105) mentioned that superficially constructing the choices seems to be relatively easy and we can produce different distractors in various ways. But if we expect testees to analyze the whole sentence in more detail and involve more worthy reading and comprehension skills, testers might not intentionally to eliminate or guess the choice and give proper reasons and understanding to each answer. However, the easiest and the most mechanic way is to increase the number of distractors, e.g. to 5-choice instead of 4-choice, which could decrease the probability of correctness simply by guessing or elimination.

Is it necessary to test grammatical knowledge independently? The place of grammar testing of foreign languages is as controversial as the place of grammar teaching. However, the target language learning differs from one's mother tongue, which decides the necessary position for a test to include grammar and vocabulary. But how to reflect this part into the TOEFL tests to maintain its quality as a communicative testing is a very important issue. One way as Rea-Dickins (1997, p. 93) mentioned, it is unnecessary to test grammar as distinct form but reflects it in some skill-based tests such as reading and writing; and it could be also in another way, which grammar should

be tested in integrative way rather than simply be put into the limited items in decontextualised single sentences as the TOEFL does.

5. Conclusion

In this paper from the aspect of grammatical knowledge, the author have attempted to examine the holistic relationship between grammar and vocabulary in the TOEFL tests and English language skills, and also the specific section, “structure and written expression”, designing features as well. From the analysis on one TOEFL test paper, it appears that the TOEFL has its fully-developed merits since it exists for quite a long time as a systematic test worldwide with high reliability, although some of the demerits make it seem not so perfect to demonstrate its validity and communicative purposes. “Knowledge of the second language is a necessary but not sufficient condition for success on the test tasks”, since success needs to be measured “in terms of performance on the task but not only in terms of knowledge of language” (Hamilton, et al., 1993, p. 350). “Structure and written expression” section attempts to use an indirect way to exam the competence of testers in writing, but it neglects its general characteristics as a communicative test in an academic environment. Because of its limited types of items with very limited related grammatical knowledge and sentence-based structure in this section, there is a suspicion about its construct validity and the role of grammar in language use in TOEFL needs to be changed. “Use of English” section in Cambridge First Certificate in English could be a good example to assess grammar if “structure and written expression” section would still be kept in the TOEFL test.

Nevertheless, diachronically the TOEFL test has been always doing self-revising and self-improving work. The TOEFL 2000 project is a broad effort under which language testing at Educational Testing Service (ETS) will evolve into the 21st century (Jamieson, et al., 2000, p. 1). It will revise the Test of Spoken English and introduce a computer-based version of the TOEFL test. However, the greatest change happened on the TOEFL 2000 test will not only depend on multiple-choice tasks but will include open-ended and constructed-response tasks as well (Jamieson, et al., 2000, p. 13). Compared the grammar and vocabulary items decades year ago with the “structure and written expression” now, necessary and evident efforts have been demonstrated and now it still continues. In the TOEFL 2000 framework, a statement indicates a tendency in the later test paper, that is, the TOEFL test may not continue to include a separate measure of structure (Jamieson, et al., 2000, p. 11), which implies that integrate the grammatical knowledge into four language skills could be an appropriate choice to meet the requirements of a communicative testing. All in all, the TOEFL test as a mean of gathering information about the EFL testers is to be used in making mainly educational decisions worldwide. No test design could be called perfect but we could find the TOEFL is trying to meet the principles of designing a test in large-scale. Like the comments from Stevenson (1987, p. 81):

> Given its purposes, examinee populations, and multiple uses and considering the attendant limitations on test content, tasks, and predictive specificity, TOEFL remains the best of its breed. Beyond those practical limitations that are necessary to its purposes and scope, TOEFL’s weakness largely reflects the state of the language testing art.

References:

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice.* Oxford University Press.

Banerjee, J. & Clapham, C. (2003). Test review: The TOEFL CBT (Computer-based Test). *Language Testing, 20*(1), 111-123.

Brown, J. D. & Ross, J. A. (1996). Decision dependability of subtests, tests and the overall TOEFL test battery. In: Michael Milanovic & Nick Saville. (Eds.). *Performance testing, cognition and assessment: Selected papers from the 15th language*

testing research colloquium, Cambridge and Arnhem. Cambridge Universtiy Press.

Carroll, B. J. (1980). *Testing communicative performance: An interim study.* Pergamon Press.

Cumming, A. & Mellow, D. (1995). An investigation into the validity of written indicators of second language proficiency. In: Alister Cumming & Richard Berwick. (Eds.). *Validation in language testing*. Multilingual Matters Ltd.

Davies, A. (1990). *Principles of language testing.* UK: Basil Blackwell.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). New Jersey: Prentice Hall.

Hamilton, J. (1993). Rating scales and native speaker performance on a communicatively oriented EAP test. *Language Testing, 10*(3), 337-353.

Hilke, R. & Wadden, P. (1997). The TOEFL and its imitations: Analyzing the TOEFL and evaluating TOEFL-prep texts. *RELC Journal, 28*(1), 28-53.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press

Huhta, A. & Randell, E. (1995). Multiple-choice summary: A measure of text comprehension. In: Alister Cumming & Richard Berwick. (Eds.). *Validation in language testing*. Multilingual Matters Ltd.

Jamieson, J., et al. (2000), *TOEFL 2000 framework: A working paper*. Educational Testing Service

Morrow, K. (1986). The evaluation of tests of communicative performance. In: Matthew Portal. (Ed.). *Innovations in language testing.* NFER-NELSON.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats (2nd ed.)*. Kluwer Academic Publishers.

Rea-Dickins, P. M. (1991). What makes a grammar test communicative? In: Charles Alderson & Brian North. (Eds.). *Language testing in the 1990s.* Macmillan Publishers Limited.

Rea-Dickins, P. (1997). The testing of grammar in a second language. In: Caroline Clapham & David Corson. (Eds.). *Encyclopedia of language and education: Language testing and assessment* (vol. 7). Kluwer Academic Publishers.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355-371.

Read, J. (2000). *Assessing vocabulary.* Cambridge University Press.

Read, J. & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1-32,

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Wadsworth Publishing Company.

Schrutt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing, 16*(2), 189-216.

Skehan P. (1991). Progress in language testing: The 1990s. In: Charles Alderson & Brian North. (Eds.). *Language testing in the 1990s.* Macmillan Publishers Limited.

Stevenson, D. K. (1987). Test of English as a foreign language. In: Alderson, J. C., et al. (Eds). *Reviews of English language proficiency tests*. Teachers of English to Speakers of Other Languages.

Vollmer, H. J. (1981). Why are we interested in general language proficiency? In: Charles Alderson & Arthur Hughes. (Eds.). *ELT documents 111- Issues in language testing.* The British Council.

Wall, D., et al. (1991). Validating tests in difficult circumstances. In: Charles Alderson & Brian North. (Eds.). *Language testing in the 1990s*. Macmillan Publishers.

Weir, C. J. (1990). *Communicative language testing.* UK: Prentice Hall.

Wijgh, I. F. (1995). A communicative test in analysis: Strategies in reading authentic texts. In: Alister Cumming & Richard Berwick. (Eds.). *Validation in language testing*. Multilingual Matters Ltd.

(Edited by Lily and Lee)